

LandScan: A Natural Language and Computer Vision System for Analyzing Aerial Images

Ruzena Bajcsy
Aravind Joshi
Eric Krotkov
Amy Zwarico

CIS Dept/D2
University of Pennsylvania
Philadelphia, Pennsylvania 19104

Abstract

LandScan (LANguage Driven SCene ANalysis) is presented as an integrated vision system which covers most levels of both vision and natural language processing. Computations are both data-driven and query-driven. In the report we focus on the design of the vision and control modules. Future work will investigate in more detail the design of the natural language interface. The data-driven system employs active control of stereo cameras for image acquisition, and dynamically constructs a surface model from multiple aerial views of an urban scene. The query-driven system allows the user's natural language queries to focus analysis to pertinent regions of the scene. This is different than many image understanding systems which present a symbolic description of the entire scene regardless of what portions of that picture are actually of interest.

1. Introduction

The aim of our research on LandScan (LANguage Driven SCene ANalysis) is to develop a system capable of dynamically updating and maintaining a model of an urban world over multiple aerial views. The system will have a natural language front end through which users can query the system about a scene, and interactively assist the vision processing by restricting the analysis to those areas of the scene which are of current interest. A unique contribution of the work is that processing is both data-driven (bottom up, determined by sensor data) and query-driven (top down, determined by user queries). The integration of both methods into one system can help overcome the shortcomings of each method employed independently. For example, if data-driven processing were able to segment a graph of edges derived from the image into several different connected components, query-driven information about what the system *should* be looking for can help impose structure, and a unique segmentation, upon the otherwise ambiguous data.

The data-driven processing starts with stereo aerial images and reconstructs the surfaces in the scene. The query-driven processing constructs a logical representation of the scene using the queries to guide analysis. High-level scene analysis is performed using an Augmented Transition Network (ATN).

This research was supported by the following grants: ARO DAA6-29-84-k-0061 AFOSR 82-NM-299 NSF MCS-8219196-CER NSF MCS 82-07294 AVRO DAAB07-84-K-F077 NIH 1-RO1-HL-29985-01.

As an example, suppose the user asks, "Is there a car on the street?" The output from this query would be: the objects to be recognized, car and street; the relation ON which must hold between them; and an indication that this query is responded to by a yes/no answer with some explanation. The vision system would then be called to find a car and a street in the relation ON. The car and street would then be added to the Scene Model (if not there already) and the system would reply with an affirmative response.

This paper will describe some related research, the implementation of the data-driven and query-driven portions of the LandScan system, and our plans for future work. A later paper will detail how natural language queries will interface with LandScan to guide the scene analysis.

2. Related Research

A large corpus of research on aerial image understanding *per se* exists, and many general vision techniques are applicable to the aerial domain. Large aerial projects have been undertaken at USC [Nevatia 83], CMU [Herman 83] and SRI [Fischler 83]. However, very few integrated systems have been successfully implemented, and the best system architecture is still an open question. In particular, the problem of providing high-level feedback to the vision system has not been adequately addressed.

ATN's have been used primarily in the domain of natural language [Bates 81], [Winograd 83]. A notable exception is the system designed by Tropf and Walter [Tropf 83] which uses an ATN model for the recognition of 3D objects with known geometries.

The work of Talmy [Talmy 83] and Herskovits [Herskovits 82] influenced the design of both the topological relations in the models and the choice of linguistic attributes which must be associated with objects in order to ensure a robust and reliable natural language interface.

3. Data-driven System Implementation and Results

This section will describe the data-driven vision modules, which must be effective in an urban world, seen from above. Urban scenes are characterized by an abundance of straight lines and planar surfaces. Under

these constraints, the scene may usefully be approximated as polyhedra.

We have tested the modules on real, highly complex aerial images; it is very difficult to present these results. We show results derived from imaging a subset of the scale model depicted in Figure 6-2: a "mock up" of an urban scene. One advantage of using the scale model is the clarity of the results, and the ease of verifying their compatibility with reality.

3.1. Vision Modules

A stereo pair of images is acquired. The gradient ∇f of the images blurred at multiple resolutions is computed, and the Canny operator is used to locally suppress non-maxima in the gradient magnitude $\|\nabla f\| = \text{SQRT}(f_x^2 + f_y^2)$. We call the surviving pixels "edgels". To find corners, the variance σ^2 in the gradient directions $\theta = \tan^{-1}(f_y)/(f_x)$ is computed over a local neighborhood. The corneriness of an edgel is proportional to the product $\sigma^2 \|\nabla f\|$.

Corresponding edgels in the two images are matched using 2-sided correlation at multiple scales. In images of parts of the scene in Figure 5-1, 83% of the vertically oriented edgels are matched, and the disparity at each is computed. The resultant sparse depth map is refined by linear interpolation (acceptable under the constraints of this domain), first across columns and then across rows. Figure 5-2 depicts the interpolated depth map (corresponding to only the more distant objects in Figure -1).

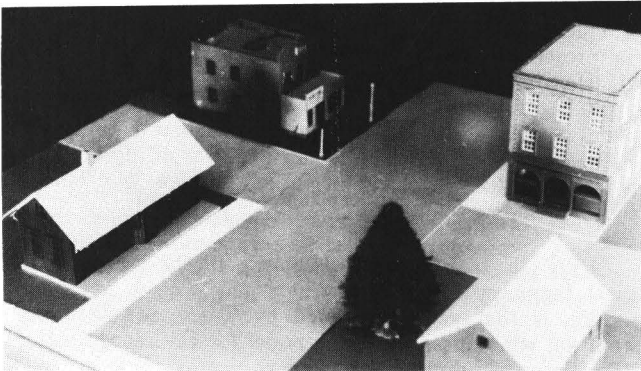


Figure 5-1: Scale model.

3.2. Surface Model

A graph is constructed to serve as the surface model [Krotkov 84]. The construction algorithm converts a set of contours into a set of closed contours represented as a graph (a linked list of vertices, edges, and faces) by traversing edges and at trihedral junctions choosing the path making the most acute angle with respect to the present path.

Surface attributes and relations are computed in the

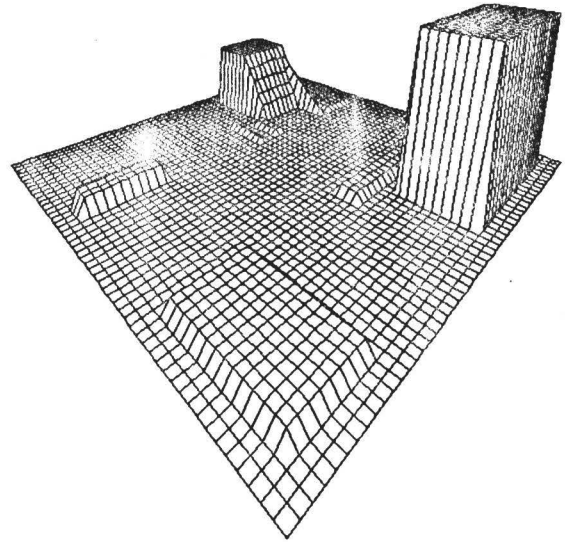


Figure 5-2: Depth map of partial view of 5-

SurfsUP [Radack, et al 84] geometrical modelling system. Attribute values for each face in the surface graph are computed: compactness, centroid, normal, area, type (building, sidewalk, field, street, and unknown), and number of sides. These values are computed once and stored on an attribute list. Computed topological relations are *above*, *adjacent* (touching), *contiguous* (sharing an edge), *contains* (proper inclusion), *looksadjacent*, *lookscontiguous* (respectively adjacent and contiguous under perspective transformations). Relations (and indirectly their complements) are computed once and stored as Boolean arrays.

4. Query-driven System Implementation and Results

This section describes the design and implementation of the query-driven processes. These include object recognition and scene modelling [Zwarico 84], high level reasoning processes, and query handling.

4.1. Object Recognition

The ATN formalism has been chosen as the paradigm for object recognition in LandScan. It is composed of three parts: the grammar, a dictionary, and an interpreter. The grammar represents the *a priori* or world knowledge that the system must have in order to recognize objects. The dictionary represents the actual data: a list of all of the faces which have been segmented by vision and the relations between them. The third component is a Lisp program which provides the control structure for the process. Figure 5-3 shows the results of running the recognizer on the scene in Figure 5-1.

4.2. The Scene Model

The Scene Model is composed of two components: a list of objects currently known to be in the scene and a set of matrices representing the primitive relations. The objects on the object list have already been recognized. Each object has associated with it a list of surfaces, its location, and a subtype. The relations which are the same as in the surface model, are represented by their adjacency matrices because the adjacency matrix is easily updated and makes composition of relations a simple matter of Boolean matrix multiplication.

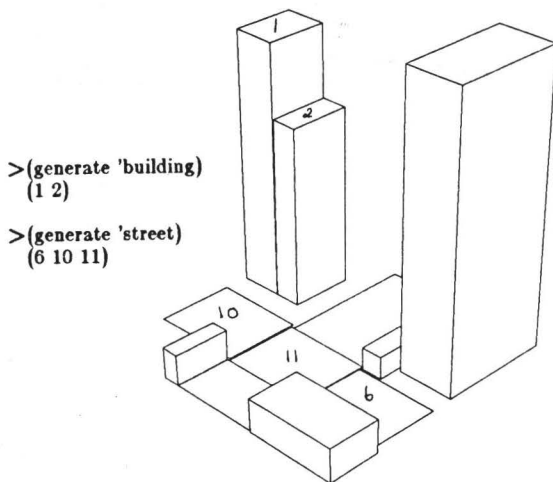


Figure 5-3: Reconstructed planar surfaces.

The Scene Model is dynamic because information can be added to it as further image analysis occurs. A new object is added to the head of the list. The relations are updated by calculating the relations between the new entity and the current object list.

4.3. Linguistic Analyzer

Given a query, the Linguistic Analyzer will symbolically represent this utterance so that it can be used by the reasoning process to analyze the image. The Linguistic Analyzer will parse the query, determine the query type, and categorize all implicit subqueries in the actual utterance. Its output will contain a list of the objects to be found, the relations which must hold between these objects, and the query type (so that an appropriate response can be generated). This is not yet implemented.

4.4. Reasoning

The Reasoner analyzes the query, determines the strategy for obtaining an answer to the query, and provides feedback to the vision system. In order to obtain the information necessary for the generation of the response the Reasoner must have both runtime data (the current Scene Model and the query) and global knowledge (the World Model and the Object Model).

The World Model describes the features and relations of the objects in the domain, buildings, streets, sidewalks, etc. The world is represented by a labelled directed multigraph in which the nodes are the objects in the domain and the arcs are labelled with the relation which can hold between them. The Object Model represents the expected physical features (subparts) and linguistic properties (features which affect the usage and interpretation of a spatial construct) of the objects in the domain.

If the reasoning processes fail to produce a positive response (the query fails to have an answer), the Reasoner performs two types of query failure analysis. The first type of query failure involves a query violating the global knowledge. In this case, the system will respond with a message indicating why the query is conceptually ill-formed in this domain. The other type of failure involves not finding the information requested in the scene model. In this case, rather than simply responding "not present",

the system may ask the user whether a new view of the scene should be analyzed.

5. Discussion

This paper has presented LandScan, a prototype integrated system under development that covers most of the different levels of vision and natural language processing. It may be used both to guide the low-level vision processing, and to provide communication of visual information to a user. While LandScan is not complete in the sense that all of it is successfully implemented, it provides a computational model for a vision system guided by natural language.

In summary, the data-driven subsystem of LandScan takes stereo images and builds a surface graph representing three-dimensional geometric and topological attributes. The query-driven modules recognize objects and build a Scene Model which represents the user's interest in the image.

The natural language interface which uses the scene representation still has to be designed. It must be able to apply locative linguistic constructs to some representation of visual data and reason about this data. When this is operative, the scene analysis will be truly query-driven and the goals of the system will have been reached.

References

- [Bates 81] Bates, Madeleine. *The Theory and Practice of Augmented Transition Network Grammars*. In Leonard Bolc (editor), *Natural Language Communication with Computers*. Springer-Verlag, 1981.
- [Fischler 83] Martin Fischler. *Image Understanding Research and its Application to Cartography and Computer-Based Analysis of Aerial Imagery*. Technical Report, SRI International, September, 1983.
- [Herman 83] Herman, Martin, Takeo Kanade, Shigeru Kuroe. *The 3D MOSAIC Scene Understanding System*. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*. 1983.
- [Herskovits 82] Herskovits, Annette. *Space and the Prepositions in English: Regularities and Irregularities in a Complex Domain*. PhD thesis, Department of Linguistics, Stanford University, 1982.
- [Krotkov 84] Krotkov, Eric. *Construction of a Three Dimensional Surface Model*. Technical Report MS-CIS-84-40, CIS Department, University of Pennsylvania, 1984.
- [Talmy 83] Talmy, Leonard. *How Language Structures Space*. Technical Report 4, Berkeley Cognitive Science Report, January, 1983.
- [Tropf 83] Tropf and Walters. *An ATN for 3-D Recognition of Solids in Single Images*. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*. 1983.
- [Winograd 83] Winograd, Terry. *Language as a Cognitive Process*. Addison-Wesley Publishing Co., 1983.
- [Zwarico 84] Amy Zwarico. *The Recognition and Representation of 3D Images for A Natural Language Driven Scene Analyzer*. Technical Report MS-CIS-84-29, University of Pennsylvania, 1984.